Making the invisible enemy visible

Structural biology plays a crucial role in the fight against COVID-19, permitting us to 'see' and understand the SARS-CoV-2 virus. However, the macromolecular structures of SARS-CoV-2 proteins that were solved at great speed and urgency can contain errors that may hinder drug design. The Coronavirus Structural Task Force has been working behind the scenes to evaluate and improve these structures, making the results freely available at insidecorona.net.

When the COVID-19 pandemic hit in early 2020, the structural biology community quickly swung into action to determine the atomic structures of the 28 viral proteins encoded by SARS-CoV-2[1]. 1392 structures covering 18 SARS-CoV-1 and SARS-CoV-2 (which comprise the subgenus *sarbecovirus*) proteins have been released over just 12 months and are freely and publicly available in the World Wide Protein Data Bank (wwPDB), which celebrates its 50th anniversary this month. These models serve as the basis for structure-based drug design and vaccine development. They are also essential for understanding how the virus hijacks human cells and causes disease. However, errors occur in even the most carefully determined structures and may be still more common in structures solved quickly and under immense pressure. Yet, even small errors can have severe consequences for structure-based drug discovery, structural bioinformatics, and computational chemistry, because they can be misinterpreted as biologically and pharmaceutically relevant.

While the wwPDB is an invaluable tool as structural biology's archive of record, it is also near-static. Released structures can only be updated by the original depositors, but there is often little motivation to make corrections once associated papers are published. 99% of PDB structure downloads are not by experimental structural biologists *per se*, but by scientists who use the structural data[2] and who might lack the training to identify and/or correct erroneous sites in the molecular model.

In this global crisis, it is vital to ensure that the available structural data are the best they can be by pushing our methods to the limit. The Coronavirus Structural Task Force, a diverse international team of structural biologists involved in methods development, responded to this challenge by rapidly categorizing, evaluating, and reviewing all SARS-CoV-1 and SARS-CoV-2 experimental structures. We do a weekly automatic post-analysis as well as a manual re-processing and re-modelling of representative structures from each of the 18 structurally characterized *sarbecovirus* proteins. Every Wednesday, when new PDB structures are released, our automatic pipeline identifies new coronavirus structures and assesses the quality of models and experimental data. This assessment, along with the original structures, is immediately made available in our online repository (**insidecorona.net**). There, we also supply a summary, an SQL database of key statistics and quality indicators, and individual results. After our validation effort began we were approached by researchers involved in *in-silico* drug screening from Folding@Home[3], OpenPandemics[4], and the EU Joint European Disruptive Initiative (JEDI). These groups aim to simulate the conformational flexibility and interaction of coronavirus proteins with each other and host cell proteins and to design small-molecule inhibitors against key SARS-CoV-2 targets via high-throughput computational modelling, a task that is exquisitely sensitive to the quality of the input model.

In addition to structure evaluation and improvement, insidecorona.net supplies literature reviews centered on the structural aspects of the viral infection cycle, host interaction partners, illustrations, and advice on selecting the best starting models for *in silico* projects. Furthermore, we have added SARS-CoV-2 proteins to Proteopedia[5] and molssi.org, as well as a 3D-Bionotes[6] deep-link into our database. Finally, we have tried to make SARS-CoV-2-related research accessible to the general public with blog posts aimed at non-scientists. We also live streamed data processing on Twitch and provided an accurate 3D printed model of SARS-CoV-2 based on deposited structures along with the files and instructions necessary to print these models.

Automatic evaluation

All macromolecular structures from SARS-CoV-1 and SARS-CoV-2 in the wwPDB are downloaded into our repository and assessed automatically within 24 hours of release. We combine new validation tools with previously developed methods, many of which were adapted for our purposes.

Crystallographic data and structure solutions

73% of reported *sarbecovirus* structures are derived by X-ray crystallography. These datasets are evaluated for pathologies including twinning, multiple lattice diffraction, ice crystal contamination, incompleteness, and radiation damage using phenix.xtriage[7] and AUSPEX[8]. Although these issues cannot be resolved after data collection, taking them into account during data processing and structure solution can yield better models. It can be difficult to identify these problems using deposited structure factors (the end result of processing raw diffraction data), since information is lost in the process. Raw data allow a more complete analysis of the experiment and re-processing - but can be difficult to obtain, as they are not deposited in the wwPDB or required for publication. We therefore invite authors to send us their raw experimental data and offer to deposit them in public repositories, such as SBGrid[9] or proteindiffraction.org[10]. All data sets we have analyzed to date have an acceptable signal-to-noise ratio; we evaluated other statistical quality indicators, examples of which are summarized in Table 1.

A general indication of how well the atomic model fits the measurement data is given by the R values: While only two structures in the our present alarmingly high R_{free} values above 35%, this does not necessarily mean an absence of modelling problems. Large R_{free} drops indicated major issues with PDB entries, especially for older SARS-CoV-1 structures. PDB-REDO[11] re-refinements generally improved R_{free}. Nevertheless, the resulting models should not be viewed as "more correct" purely on the basis of a lower R value, particularly at lower resolution where the relationship between R values and model quality degrades[12]. Critical manual inspection of the model remains necessary.

Structures from single-particle Cryo-EM

Cryo-EM structures make up 24% of reported SARS-CoV-1 and SARS-CoV-2 structures. Raw data are not available from the wwPDB, but deposition into EMPIAR[13] is increasingly common. The reconstructed 3D map deposited in the EMDB[14] allows calculation of the fit between model and map by Fourier Shell Correlation (FSC) to assess agreement between features at different resolutions. FSCs, real-space Cross-Correlation Coefficient (CCC), Mutual Information (MI) and Segment Manders' Overlap Coefficient (SMOC)[15] were calculated with the CCP-EM[16] model validation task (see Table 1). While MI and CCC are single value scores that indicate how well model and map agree overall, the SMOC score evaluates the fit of each modelled residue individually and can highlight specific regions where model and map disagree. We use Haruspex[17], a neural network trained to recognize secondary structure elements and RNA/DNA in cryo-EM maps, as visual guidance for manual structure evaluation.

Evaluation of the structural models based on prior knowledge

MolProbity[18] is used to evaluate the model quality, check covalent geometry and conformational parameters of protein and RNA, and steric clashes. Some of these traditional quality indicators are used as additional restraints during refinement, which reduces their usefulness as quality metrics. The newer MolProbity CaBLAM score[6] is designed to find local errors and particularly useful at 3-4 Å resolution. Current refinement packages do not specifically aim at improving this score, arguably making it a more reliable quality indicator. In addition, checking the amino acid sequence of each model against the one specified in the deposited PDB file highlighted mismatches in 23 cases.

During the COVID19 crisis the MolProbity webservice has been pushed to its limit as drug developers screen the same SARS-CoV-2 structures many times. We developed a custom MolProbity pipeline which makes the validation results for these structures available online, thereby decreasing the webservice's workload.

Manual evaluation

Although the structural biology community has achieved a high level of automation in data collection, data processing and structure solution in recent years, the process of structure determination still requires interpretation by researchers. This especially applies to low-quality maps with poor fit between experimental data and structural models. Visual residue-by-residue inspection by an experienced structural biologist remains the best way to judge quality. We therefore select representative structures of each SARS-CoV-2 protein, as well as those of particular interest for drug development for manual evaluation. Certain problems are surprisingly common, such as peptide bond flips (Fig. 1C, 1D), rotamer errors, occupancy problems (Fig. 1E) and misidentification of small molecules or ions, e.g. water as magnesium and chloride as zinc. Of note, zinc plays an important role in many SARS-CoV-2 proteins. We found many zinc coordination sites to be mismodelled, with the zinc ion missing or pushed out of density and/or erroneous disulphide bonds between the coordinating cysteine residues (Fig. 1A, 1B, 1H). In addition, many coronavirus proteins are glycosylated at surface asparagine residues, but glycan sugars were often flipped from their correct orientation around the N-glycosidic bond (Fig. 1F, 1G). This can be avoided by using tools such as Privateer[19] and the automated carbohydrate building tool in Coot[20]. It is important to note that deviation from expected behavior is not always an error and can also be a functionally relevant feature, e.g. the strained geometries often found at catalytic sites. However, such deviations must be strongly supported by the experimental data.

Of the structures we checked manually, we were able to substantially improve 31 in terms of model quality, data quality, or both. Here we give two examples to illustrate the importance of careful inspection of experimental data and resulting models.

Papain-like protease

Nsp3 (SARS-CoV-2 Non-structural-protein 3) contains a papain-like protease domain that is essential for infection because it cleaves the viral polypeptide. The first SARS-CoV-2 structure (PDB 6W9C) was released 1st April 2020, only three months after the viral genome (GenBank: MN908947.2) was reported[21], and it was immediately used in drug design efforts. The overall completeness of the measured data, however, was only 57%. Examination of the raw data, available from proteindiffraction.org[10], revealed strong radiation damage, exacerbated by a poor data collection strategy. This could not be deduced from the PDB deposition, underlining the importance of the availability of raw data.

The crystal has 3-fold non-crystallographic symmetry with each papain-like protease domain monomer containing a functionally important Zn^{2+} ion bound by four cysteines with similar $C_ß$-$S_γ$-Zn angles and Zn- $S_γ$ bond lengths. Because of radiation damage, these sites have poor density. One site has been modelled as a disulphide bond and two free cysteines (Fig. 1H) while the other two coordinated zinc with strongly varying $C_ß$- $S_γ$-Zn angles and Zn-S bond lengths.

We reprocessed the images using XDS[22], a software for the processing of single-crystal X-ray diffraction images. The Staraniso server (http://staraniso.globalphasing.org) was used to determine and apply an anisotropic limit for the diffraction data. This careful manual intervention improved overall data quality and resolution from 2.7 to 2.6 Å, but the revised overall ellipsoidal completeness was only 44.5%. Adding zincs to all sites, restraining the bond lengths and angles to the expected values, using NCS (non-crystallographic symmetry) restraints and an overall higher weighting of ideal geometry, together with remodeling of side chains and water molecules improved the electron density maps and lowered the R values by 4%. This exemplifies the interconnection between data collection, data processing and model building: even if the data collection strategy is not ideal, taking the resulting problems into account during data processing and refinement can drastically improve the final model.

A structure of the C111S mutant of Papain-like protease domain (PDB 6WRH) was released a month later. In this structure, the zinc sites were clearly resolved in all subunits. In the meantime, however, PDB 6W9C had been widely used in *in silico* drug design. 20% of the over 140 research teams in the JEDI COVID19 GrandChallenge, a competition to find potential drugs against COVID-19 *in silico*, have used this model. The availability of a better structure a month earlier would have increased their chances of success and saved computing and person hours.

RNA polymerase complex

SARS-CoV-2 replicates its single-stranded RNA genome using a macromolecular complex of RNA-dependent RNA polymerase (Nsp12; RdRp), Nsp7 and Nsp8 (SARS-CoV-2 Non-structural-proteins 7 and 8, respectively). Earlier Cryo-EM structures of the SARS-CoV-1 homologs (PDB 6NUR, PDB 6NUS) include a disordered unmodeled loop followed by a visible but short and irregular helix and a flexible C-terminus. Density for this helix was poorly resolved, but the model had valid geometry. Our analysis of one of the first structures of the equivalent SARS-CoV-2 complex (PDB 7BTF) revealed that the sequence in this C-terminal region (part of the RNA binding groove) was misaligned by nine residues (Fig. 2). This error was present in all related SARS-CoV-1 and SARS-CoV-2 structures, likely because new structure determination typically starts from an earlier model if one is available.

A structure of the RdRp complex bound to the nucleotide analog remdesivir (PDB 7BV2[23]) was released soon after and provided the basis for rational design of related drug candidates[24]. This structure also featured the 9-residue sequence misalignment. We rebuilt the structure using ISOLDE, CaBLAM, and visual inspection, correcting some flipped or *cis*-vs-*trans* peptides (see Fig 1C, 1D) and three RNA conformers near remdesivir, including a backward adenosine base, and were able to add several residues and waters with good density and geometry. Remdesivir is covalently attached to the RNA, but it is only present in an estimated ≤ 50% of the measured molecules[12]. This means the active site is a mixture of at least two different states, so unsurprisingly the modeled Mg^{2+} ions and pyrophosphate are poorly supported by the experimental density and local contacts. This is of concern for subsequent *in-silico* docking and drug design, which often take all atoms in the deposited structure as a fixed framework to build into. The remodeled structures of the complex might offer a more solid basis for drug design even if the half-occupancy active site was not widely discussed[12]. It is notable that despite the very large register error and various smaller issues, by traditional "summary" metrics the model appeared extremely good, with no Ramachandran nor rotamer outliers and a clash score of 2, highlighting that direct visual inspection must remain a key step in any modelling process.

Although the problems discussed above were present in the originally deposited structures, nearly all are now corrected. This was achieved by making corrected models available on our website and contacting the original authors of these structures with detailed descriptions, allowing them to deposit revised versions to the wwPDB at their discretion.

In conclusion

In the last 40 years, structural biology has achieved a high level of automation, and methods have advanced greatly. It is now feasible to solve a new structure from start to finish in a matter of weeks with little specialist knowledge. This is exemplified by the rapid solution of SARS-CoV-2 structures during the pandemic, which is a remarkable achievement. These structures have enabled rapid progress in the development of therapeutics and vaccines. However, errors at all stages of structure determination are not only common, but often remain undetected. Unfortunately, no individual researcher can be fully conversant with all the details of structure determination, chemical properties of interacting groups, catalytic mechanisms, and the viral infection cycle. While any molecular model could benefit from examination by multiple experts, it is particularly important to rapidly carry out such inspection of Coronavirus-related structures.

Structural models are an interpretation of the measured data, and deposited structures could be seen as the first round of interpretation which may provide considerable biological insight but may leave room for improvement. The availability of raw data would allow more complete assessment of the structure solution. It would also offer the opportunity to reanalyze the data and to propose updates to the original authors or to deposit derivative models in the wwPDB. We believe that, as a community, we need to change how we see, address, and document errors in structures to achieve the best possible structures from our experiments. We are scientists: *In the end, truth should always win.*

Acknowledgements
This work was supported by the German Federal Ministry of Education and Research [grant no. 05K19WWA], Deutsche Forschungsgemeinschaft [project TH2135/2-1], the Wellcome Trust [grants 208398/Z/17/Z and 209407/Z/17/Z], and the US National Institutes of Health [grant R35 GM131883]. It would not have been possible without exchange, discussions and support from the computational and experimental structural biology community; particularly Lu Zhang, John Chodera, Stefano Forli, Thomas Hermanns, Paul Emsley, Tom Burnley, Clemens Vonrhein, Iris Young, James Fraser and Arwen Pearson. We would also like to thank to Holger Theymann, Nicole Dörfel and Thomas Splettstößer for web design and visualization of our work. Lastly, we are grateful to Elisa Bandello, Pairoh Seeliger & Florian Platzmann for their continued support.

Competing Interests
The authors declare no competing interests.

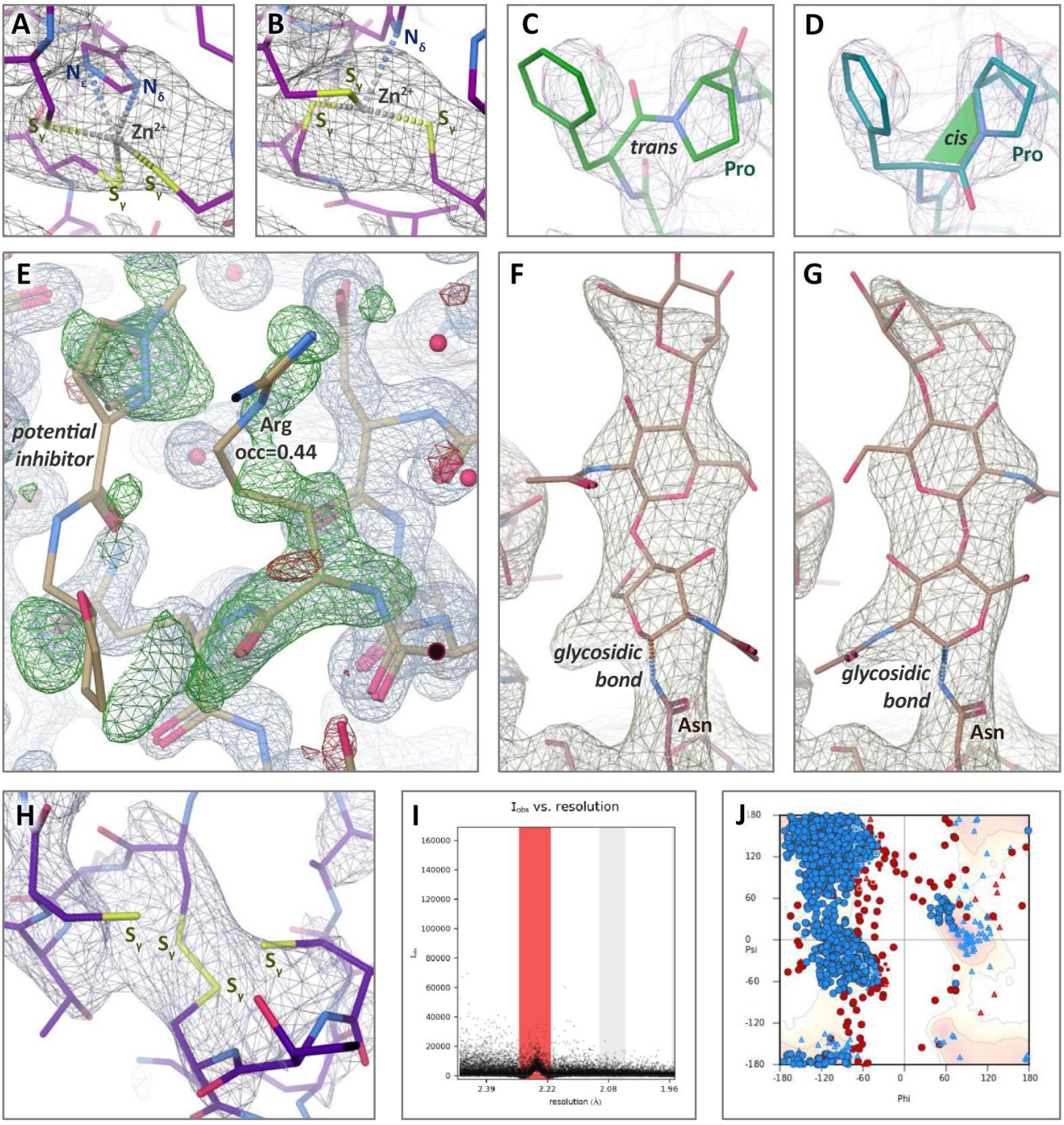

Fig. 1 Potential for improvement. *All pictures except I are screenshots from Coot 0.9.9-pre-release. Residual density and reconstructions maps in blue-gray; difference electron density in red and green.* ***A.*** *SARS coronavirus Nsp14-Nsp10 (PDB 5C8T) histidine zinc coordination site (B603) with residual density contour level 0.445, rmsd 0.150.* ***B.*** *Histidine from* ***A*** *has been swapped in ISOLDE, leading to tetrahedral coordination of Zn^{2+}, followed by PDB-REDO refinement with manually added links.* ***C.*** *Proline A505 is modelled as trans in RdRp complex (PDB 7BV2, left) but density indicates a cis main chain conformation (****D****). The deposited PDB entry was updated after we contacted the original authors.* ***E.*** *High difference electron density at residue A165 in SARS-CoV-2 main protease (PDB 5RFA) due to only 0.44 occupancy instead of 1.00 near potential inhibitor (left). Residual map contour level 0.54, rmsd 0.319, difference density at contour level 0.35, rmsd 0.114.* ***F.*** *SARS-CoV-2 spike receptor binding domain complexed with human ACE2 (PDB 6VW1): this N-linked glycan is flipped approximately 180° around the N-glycosidic bond. After we contacted the original authors, this entry was revised - see* ***G.*** *Correction improves the density fit of the sugar chain. Residual map at contour level 0.311 rmsd 0.265.* ***H.*** *Disulphide bond A226-A189 in papain-like protease (PDB 6W9C) with electron density at contour level 0.214 rmsd 0.136. While the density map does not indicate a zinc, it is a zinc finger domain; the other NCS copies have a zinc coordinated here and the other two cysteines are uncoordinated.* ***I.*** *AUSPEX[8] plot of SARS-CoV main protease (PDB 2HOB); ice rings are reflected by a bias in the intensity distribution (red).* ***J.*** *Ramachandran plot or torsion angles in the peptide bond for SARS-CoV Nsp10/Nsp14 dynamic complex (PDB 5NFY); usually, there should only be a few outliers (red) as most peptide bonds adhere to typical angular distributions. Picture: CSTF/insidecorona.net.*

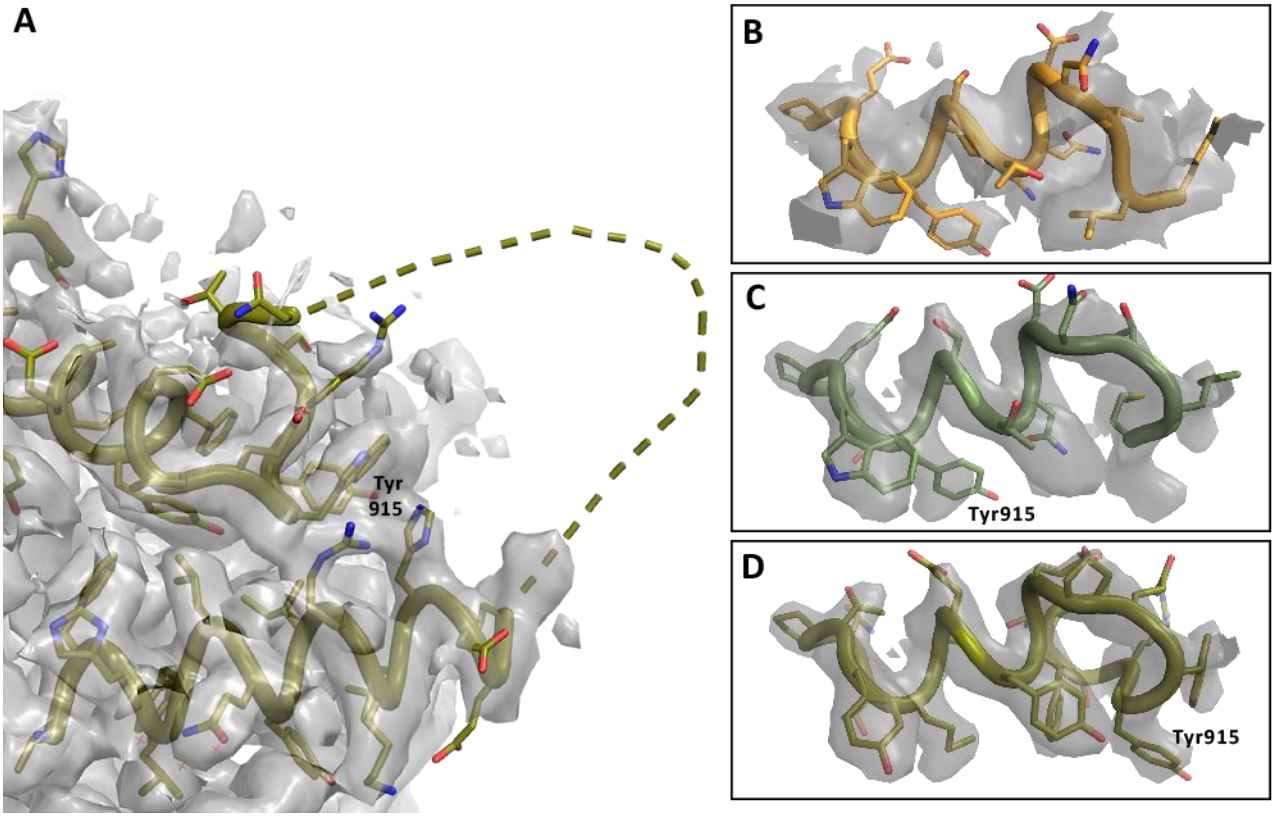

Fig. 2. Registry shift in C-terminus of RNA Polymerase. ***A.*** *Overview with missing loop shown as dashed line (PDB 7BV2); map at 2.4σ.* ***Right side:*** *Details of C-terminal helix at 5σ.* ***B.*** *Lower resolution map and model PDB 6NUS. Judging the side chain fit is difficult.* ***C.*** *Higher resolution map and model PDB 7BV2 as deposited; the side chain fit is suboptimal due to the register error.* ***D.*** *Amended model for PDB 7BV2; the side chains now fit the density. The register shift is indicated by labelled Tyr915. Picture: CSTF/insidecorona.net.*

Key indicators in evaluation	Number of depositions (% of total)
Crystallography (999 depositions)	
Completeness < 80%	14 (1.4%)
R_{free} > 35%	2 (0.2%)
Potential twinning	52 (5.2%)
Contaminated by ice diffraction	93 (9.3%)
Incorrect mask	86 (8.6%)
Single-particle Cryo-EM (360 depositions)	
Average model-map FSC < 0.4	46 (13%)
MI score < 0.4	56 (16%)
SMOC > 10%	64 (18%)
Other indicators (1392 depositions)	
CaBLAM outlier conformations > 2.0%	310 (22%)
CaBLAM severe C_α outliers > 1.0%	106 (7.6%)
Sequence mismatch	23 (1.7%)

Table 1. Examples of quality indicators pointing to potential problems in PDB entries, calculated in our automatic evaluation pipeline. Potential twinning as identified by L-Test[7], ice diffraction and incorrect mask identified by visual inspection of AUSPEX plots[8]. The chosen cutoffs for FSC and MI score[16] indicate poor overall agreement between map and model. The SMOC[15] score > 10% indicates more than 10% of the residues of a structure fit poorly with the map and could potentially be improved.

Literature

1. Baker, E. N. Visualizing an unseen enemy; mobilizing structural biology to counter COVID-19. *Acta Cryst D* **76**, 311–312 (2020).
2. Burley, S. K. *et al.* RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education. *Protein Science : A Publication of the Protein Society* **27**, 316 (2018).
3. Zimmerman, M. I. *et al.* Citizen Scientists Create an Exascale Computer to Combat COVID-19. *bioRxiv* 2020.06.27.175430 (2020) doi:10.1101/2020.06.27.175430.
4. OpenPandemics - COVID-19 | Research | World Community Grid. https://www.worldcommunitygrid.org/research/opn1/overview.do.
5. Proteopedia: A status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. *Journal of Structural Biology* **175**, 244–252 (2011).
6. Prisant, M. G., Williams, C. J., Chen, V. B., Richardson, J. S. & Richardson, D. C. New tools in MolProbity validation: CaBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics. *Protein Science* **29**, 315–329 (2020).
7. Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. 9.
8. Thorn, A. *et al.* AUSPEX: a graphical tool for X-ray diffraction data analysis. *Acta Cryst D* **73**, 729–737 (2017).
9. Morin, A. *et al.* Cutting edge: Collaboration gets the most out of software. *eLife* **2**, (2013).
10. Grabowski, M. *et al.* A public database of macromolecular diffraction experiments. *Acta Cryst D* **72**, 1181–1193 (2016).
11. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB_REDO server for macromolecular structure model optimization. *IUCrJ* **1**, 213–220 (2014).
12. Croll, T. I., Williams, C. J., Chen, V. B., Richardson, D. C. & Richardson, J. S. Improving SARS-CoV-2 structures: Peer review by early coordinate release. *Biophysical Journal* **120**, 1085–1096 (2021).
13. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat Methods* **13**, 387–388 (2016).

14. Lawson, C. L. *et al.* EMDataBank.org: unified data resource for CryoEM. *Nucleic Acids Res* **39**, D456–D464 (2011).

15. Joseph, A. P. *et al.* Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100**, 42–49 (2016).

16. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Cryst D* **73**, 469–477 (2017).

17. Mostosi, P., Schindelin, H., Kollmannsberger, P. & Thorn, A. Haruspex: A Neural Network for the Automatic Identification of Oligonucleotides and Protein Secondary Structure in Cryo-Electron Microscopy Maps. *Angewandte Chemie International Edition* (2020) doi:10.1002/anie.202000421.

18. Williams, C. J. *et al.* MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science : A Publication of the Protein Society* **27**, 293 (2018).

19. Agirre, J. *et al.* Privateer: software for the conformational validation of carbohydrate structures. *Nat Struct Mol Biol* **22**, 833–834 (2015).

20. Emsley, P. & Crispin, M. Structural analysis of glycoproteins: building N-linked glycans with Coot. *Acta Cryst D* **74**, 256–263 (2018).

21. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

22. Kabsch, W. XDS. *Acta Cryst D* **66**, 125–132 (2010).

23. Yin, W. *et al.* Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* **368**, 1499–1504 (2020).

24. Zhang, L. *et al.* 1’-Ribose cyano substitution allows Remdesivir to effectively inhibit nucleotide addition and proofreading during SARS-CoV-2 viral RNA replication. *Phys. Chem. Chem. Phys.* **23**, 5852–5863 (2021).